# KI-Methoden für die automatische Anreicherung von räumlich-semantischen Tatortmodellen

Steffen Franz[1], Timo Bittner[2], Robert Irmler[1], and Christian Eller[2]

[1]Kriminaltechnisches Institut · Bundeskriminalamt · Äppelallee 45 · 65203 Wiesbaden · E-Mail: KT24-INSITU@bka.bund.de
[2]Institut für Numerische Methoden und Informatik im Bauwesen · Technische Universität Darmstadt · Franziska-Braun-Str. 7 · 64287 Darmstadt · E-Mail: insitu@iib.tu-darmstadt.de



Das Sicherheitsforschungsprojekt INSITU beschäftigt sich mit der Optimierung der Strafverfolgung mit Hilfe einer Mobilcomputer-basierten Vor-Ort-Beschreibung von Tatorten. Um eine Beschleunigung des zeitintensiven Prozesses der Dokumentation aller Informationen und Zusammenhänge zu erreichen, wird in diesem Beitrag demonstriert, wie ein Mobilcomputer in die Lage versetzt werden kann, Objekte und deren räumlich-semantische Beziehungen automatisiert zu erfassen. Hierbei werden Augmented-Reality-Technologien mit maschinellen Lernverfahren fusioniert und auf einem Mobilcomputer in Echtzeit zur Anwendung gebracht. Mittels AR-Technologien verortet sich das Gerät im Raum, verfolgt Positionsänderungen und erzeugt kontinuierlich Bilddaten, welche parallel über Algorithmen für die Objektdetektion mit semantischen Informationen angereichert werden. Die so generierten Daten werden anschließend mit den Tracking-Informationen vereinigt, wodurch das räumlich-semantische Tatortmodell automatisiert um Objektinformationen erweitert werden kann.

**Keywords**: Tatortdokumentation, Maschinelle Lernmethoden, Augmented Reality, Bildverarbeitung

## 1 Einleitung

Unabhängig davon, ob der Baufortschritt eines Gebäudes, der Zustand einer Immobilie oder ein Tatort dokumentiert werden sollen, es fallen vor Ort eine Vielzahl von Informationen an, die in Beziehung zueinander erfasst werden müssen. Zu den Dokumentationsinformationen gehören neben der eigentlichen Beschreibung des Objekts auch dessen Kontext (z.B. Lage, Aufnahmezeitpunkt und erfassende Person). Dabei werden die zu dokumentierenden Informationen oftmals mit unterschiedlichen Geräten, digital und analog aufgenommen. Eine Zusammenführung erfolgt in der Regel erst im Anschluss an die Aufnahme in mühsamer Handarbeit. Das Sicherheitsforschungsprojekt INSITU beschäftigt sich daher mit der Optimierung der Strafverfolgung durch eine Mobilcomputer-basierte Vor-Ort-Beschreibung von Tatorten. Ein Forschungsziel ist dabei, den Nutzer mit Hilfe handelsüblicher, preiswerter Geräte bei der Dokumentation durch Automatisierung von Aufgaben zu entlasten. Insbesondere die zuvor angesprochenen Kontextinformationen der Dokumentation lassen sich mit Hilfe eines Mobilcomputers gut (teil-)automatisiert erfassen. Dieser Beitrag soll demonstrieren, wie unter Verwendung aktueller Forschungen im Bereich der KI-Methoden eine weitestgehend maschinelle Unterstützung einer Dokumentation gestaltet werden kann. Hierfür wird ein Verfahren vorgestellt, welches eine automatische Objekterkennung, inklusive räumlicher Verortung in einem dreidimensionalen Modell, in Echtzeit vor Ort ermöglicht.

## 2 Stand der Technik

Zur Dokumentation geometrischer Daten kommen heute typischerweise stationäre und mobile Laserscanner sowie photogrammetrische Verfahren zum Einsatz. Die so aufgezeichneteten Punktwolken bilden die dokumentierte Umgebung lediglich geometrisch ab. Sie geben keine beschreibende Auskunft über aufgezeichnete Objekte oder deren Zusammenhänge. Für eine automatisierte, maschinelle Verarbeitung der Daten sind diese Informationen jedoch notwendig. Daher gibt es eine Vielzahl an Forschungen, die sich mit der (teil-)automatischen Extraktion von semantischen Informationen aus Punktwolkendaten beschäftigen (Czerniawski et al., 2016), (Beetz et al., 2016), (Qi et al., 2019). Gemein haben diese Ansätze, dass neben dem Einsatz von Spezialhardware zwischen Aufnahme und Analyse differenziert wird. Die semantisch angereicherten Informationen stehen somit nicht sofort vor Ort zur weiteren Verwendung zur Verfügung. Die simultane Aufnahme und Prozessierung der Informationen vor Ort, in Echtzeit sowie auf einem mobilen Endgerät demonstrieren Wald et al. (2018) und Franz et al. (2018). Es kommen dabei allerdings ebenfalls spezielle Mobilgeräte mit einer TOF-Kamera zum Einsatz.
Eine Voraussetzung zur geometrischen Erfassung der Umgebung mit Mobilgeräten ist, dass diese sich im dreidimensionalen Raum lokalisieren und ihre Positionsveränderung verfolgen können (Simultaneous Localization and Mapping). Im Bereich der Konsumerhardware existieren hierfür Augmented Reality (AR) Softwarebibliotheken. Zu den am weitesten verbreiteden zählen Google ARCore (Google, 2019*a*) und Apple ARKit (Apple, 2019). Beide Bibliotheken ermöglichen zusätzlich die Wiedererkennung von zweidimensionalen Bilddaten, ARKit ist zudem seit Version 2 in der Lage dreidimensionale Gegenstände wiederzuerkennen. In beiden Fällen müssen die zu erkennenden Objekte dem System vorher angelernt werden. Das System ist im Anschluss dann in der Lage genau diese wiederzufinden. Ähnliche Objekte, die sich beispielsweise in Farbe oder Form unterscheiden, werden nicht mehr erkannt. In unkontrollierten Umgebung, wie beispielsweise an Tatorten, ist es nicht möglich diese Objekte zunächst für die Wiedererkennung durch das System anzulernen.
Eine allgmeine Objekterkennung ist jedoch in Bilddaten mit Hilfe künstlicher neuronaler Netze, insbesondere Convolutional Neural Networks (CNN), möglich. Moderne Objektdetektoren - wie Faster R-CNN (Ren et al., 2017), R-FCN (Dai et al., 2016), SSD (Liu et al., 2016) und YOLO (Redmon et al., 2016) - basieren auf diesen Netzwerken. Sie sind effizient und schnell genug, um auch auf mobilen Geräten ausgeführt werden zu können. Zhang et al. (2018) und Liu et al. (2019) entwickelten cloudbasierte System für die automatische Detektion und Verortung von Objekten. Um die Schnelligkeit zu erhöhen und Aussagen in Echtzeit treffen zu können, wurde die GPU eines externen Servers genutzt und Ergebnisse an das mobile Endgerät gesendet, während dies weiter die Position verfolgt und Daten übermittelt. Objekte werden auf diese Weise beschriftet und mit geringer Latenz auf dem mobilen Endgerät angezeigt. Voraussetzung für diese Anwendung ist die sichere und performante Anbindung an die Cloud. Eine Lösung zur allgemeinen Objekterkennung und Verortung, welche autark auf einem handelsüblichen Mobilgerät ohne Anbindung an eine Cloud und spezielle Sensoren auskommt, existiert nicht. Im Rahmen dieser Arbeit wurde ein Verfahren entworfen dies zu ermöglichen.

# 3 Konzept

Das dem Forschungsprojekt INSITU zu Grunde liegende digitale Tatortmodell zeichnet sich durch eine räumlich-semantische Beschreibung aller Elemente eines Tatort aus. Dies bedeutet, dass Objekte zusätzlich zu bezeichnenden Merkmalen auch über eine Koordinate verfügen, welche die Lage des Objekts in einem lokalen dreidimensionalen kartesischen Koordinatensystem eindeutig beschreibt. Damit ein Computer in die Lage versetzt werden kann, diese Informationen automatisiert aus Bilddaten zu erfassen, muss er folgende zwei Kernprobleme lösen:

1. **Objekterkennung:** Um welches Objekt handelt es sich im Bild?
2. **Positionsbestimmung:** Wo befindet sich das Objekt in Relation zum lokalen Koordinatensystem des Tatortmodells?

Im Rahmen des Forschungsprojekts INSITU müssen diese Probleme vor Ort, während der Bildaufnahme, in Echtzeit auf einem Smartphone oder Tablet gelöst werden. Dies ist notwendig, da die extrahierten Informationen sofort dem Benutzer zur weiteren Verwertung zur Verfügung stehen sollen. Zur Lösung des ersten Problems, der Objekterkennung, wurden in den vergangen Jahren eine Vielzahl von Methoden im Bereich des maschinellen Lernens entwickelt. Zu den aktuell gängigsten Deep Learning Ansätzen gehören: Region Proposals (R-CNN, Fast R-CNN, Faster R-CNN), Single Shot MultiBox Detector (SSD) und You Only Look Once (YOLO). Diese verwenden ein künstliches neuronales Netz, dass zuvor mit einer großen Menge an Bilddaten für die Detektion bestimmter Objekte trainiert wurde. Zusätzlich zur Erkennung des Objekts geben diese Verfahren die Position des Objekts im Bild an und treffen eine Aussage darüber, wie sicher sie sich bei einem detektierten Objekt sind. Die Position wird durch eine Bounding Box gekennzeichnet, deren Koordinaten in Bezug auf das untersuchte Bild angegeben werden (siehe Abbildung 1 a)). Um die zweite Fragestellung beantworten zu können, müssen Tiefeninformationen aus dem Bild gewonnen werden. Dies ist mit zusätzlichen Sensoren (z.B. TOF-Kamera, Stereokamera) möglich, erfordert jedoch Spezialhardware. Dank der jüngsten Entwicklungen im Bereich der Augmented-Reality-Technologien stehen allerdings auch robuste Algorithmen zur Verfügung, die eine Positionsbestimmung und Kartierung der Umgebung auf Basis der Inertialsensoren in Kombination mit der Kamera eines Smartphones ermöglichen. Zu nennen sind hier Visual Inertial Odometry (VIO) sowie Simultaneous Localization and Mapping (SLAM). Die Verfolgung der Position wird dabei durch markante Stellen (“feature points”) im Bild gestützt. Aufgrund der Verfolgung dieser Punkte über mehrere direkt aufeinander folgende Bilder hinweg, können diese zur Bestimmung von Tiefeninformationen verwendet werden. In Abbildung 1 b) werden die detektierten “Feature Points” visualisiert. Durch die simultane Anwendung der beiden KI-Methoden ist ein Computer in der Lage, automatisch Objekte in einem Live-Kamerafeed vor Ort zu erkennen und zu verorten. Die Idee ist, die Ansammlung von “Feature Points” wie eine dünnbesetzte Punktwolke zu betrachten. Punkte, die im aktuellen Bild vorhanden sind, werden in den zweidimensionalen Objektraum des Bildes transformiert. Fallen sie dort in den Detektionsbereich (Bounding Box) eines Objektes, werden sie mit der Bezeichnung des Objekts gelabelt.


(a) (b)

**Abbildung 1:** Maschinelle Auswertemöglichkeiten von Bilddaten: a) Objekterkennung; b) Feature Points

Bewegt man das Gerät durch den Raum werden Objekte aus unterschiedlichen Winkeln detektiert. Punkte, die aus mehreren Blinkwinkeln mit dem gleichen Objekt gelabelt wurden, können mit einer hohen Sicherheit als zutreffende Positionsangabe für das Objekt gewertet werden. Nach Abschluss des Scans wird als geometrische Repräsentation des Objekts die Bounding Box um alle zum Objekt gehörenden Punkte gewertet. Auf diese Weise wird das räumlich-semantische Tatortmodell mit Hilfe von KI-Methoden automatisch angereichert. Das nächste Kapitel erläutert die Umsetzung.

## 4 Umsetzung

Zur Validierung des Konzepts wurde eine Android-Applikation entwickelt. Im Folgenden wird zunächst auf die wesentlichen Komponenten eingegangen, im Anschluss wird das entwickelte Verfahren zur automatischen Anreicherung des räumlich-semantischen Tatortmodells vorgestellt.

**Plattform:** Die App ist lauffähig auf allen Smartphones und Tablets, die mindestens Android SDK Version 26 (Android 8.0 Oreo) und Google ARCore unterstützen. In der vorliegenden Arbeit wurden ein Samsung Galaxy S4 Tablet (Android 8.1) sowie ein Google Pixel 2 (Android 9.0) verwendet.

**Objekterkennung:** Für die Objekterkennung wird in der vorliegenden Arbeit Tensorflow Lite in der Version 1.13.1 eingesetzt. Die Detektion wird durch ein von Google bereitgestelltes quantisiertes MobileNet-SSD-Modell (Google, 2019*b*) durchgeführt, welches mit dem COCO-Bilderdatenset (Lin et al., 2014) trainiert wurde und 80 Objekte aus dem alltäglichen Gebrauch differenzieren kann (z.B. Tisch, Laptop, Tasse). **Positionsbestimmung:** Zur Bestimmung sowohl der Lage des Geräts im Raum als auch zur Tiefenmessung wird Google ARCore (Google, 2019*a*) in der Version 1.9 verwendet. Das Gerät ist dadurch fähig seine Position und Orientierung autark zu bestimmen. Es wird keine externe Infrastruktur benötigt, lediglich ausreichend gute Lichtverhältnisse. Die zur Positionsbestimmung verwendeten "Feature Points" erhalten während einer Session eine eindeutige ID. Dies ermöglicht die Wiedererkennung von Punkten bis zur Beendigung der App.

## 4.1 Automatische Anreicherung eines räumlich-semantischen Tatortmodells

Zunächst muss das System lokalisiert werden. Dies geschieht mittels Scan eines Markers, welcher am Einsatzort angebracht wird. Dieser stellt den Ursprung des Koordinatensystems dar. Koordinaten, welche sich auf diesen Ursprung beziehen werden im Folgenden als Weltkoordinaten bezeichnet. Nach erfolgreicher Lokalisierung verfolgt das System die eigene Lageänderung und verarbeitet die durch die Hauptkamera des Geräts aufgenommen Bilddaten. Der gesamte Prozess ist in Abbildung 2 dargestellt.

**Frame aufzeichnen:** Solange das Tracking der Lageänderung gültige Werte liefert, werden die dafür verwendeten Frames (Bildgröße von 640 x 480 Pixel) aufgezeichnet und analysiert. Der Algorithmus verarbeitet die Frames sequentiell und ignoriert neu eintreffende Frames solange sich ein Frame in der Verarbeitung befindet. Bevor der Frame zur Objekterkennung weitergebeben wird, werden zusätzlich noch die Position und Orientierung sowie Projektions- und Viewmatrix der Kamera des Frames gespeichert. Weiterhin wird die Summe der "Feature Points" als lokale Punktwolke des Frames gesichert.

**Objekt(e) detektieren:** Im nächsten Schritt werden Objekte im aktuellen Frame mit Hilfe des künstlichen neuronalen Netzes detektiert. Wurde mindestens ein Objekt gefunden, wird die Liste der Objekte inklusive ihrer Positionen im Frame (Bounding Box) gespeichert und die Analyse der Punktwolke gestartet.

**Transformation auf 2D-Bildkoordinaten:** Eingangs wird geprüft, ob sich die Punkte innerhalb des Sichtfeldes der Kamera und in einem Maximalabstand von 2m zum Gerät befinden. Der Maximalabstand wurde gewählt, da mit zunehmender Entfernung die Messungenauigkeit zunimmt. Erfüllt ein Punkt nicht diese Randbedingungen, wird er als Ausreißer betrachtet und verworfen. Danach folgt die Vorbereitung auf den Abgleich der erkannten Objekte mit der Punktwolke. Hierfür müssen die Punkte zunächst aus ihrem dreidimensionalen Raum auf die 2D-Bildkoordinaten des Frames transformiert werden.

**Punkt labeln:** Der Labelingprozess beinhaltet die Prüfung, ob ein Punkt der Punktwolke innerhalb einer Bounding Box eines Objekts liegt. Ist dies der Fall, wird der Punkt mit der Bezeichnung des Objekts annotiert und in einer Liste gespeichert.

**Transformation auf Weltkoordinaten:** Eine Eigenschaft der ARCore-Bibliothek ist, dass jede Punktwolke in Bezug auf den aktuellen Frame registriert ist. Dies bedeutet, dass die Punkte auf Weltkoordinaten transformiert werden müssen. Ist dies geschehen, werden sie in einer Liste gespeichert und ein Zähler inkrementiert, sofern für einen bereits vorhandenen Punkt das gleiche Label erneut erkannt wurde. Ist das Gegenteil der Fall, wird der Zähler dekrementiert. **Objekt dokumentieren:** Wurde ein Punkt mehrfach mit dem gleichen Label registriert, ist die Wahrscheinlichkeit hoch, dass an dieser Position das entsprechende Objekt erkannt wurde. Die Information wird in einer Liste gespeichert und sofern das Tracking noch aktiv ist, beginnt der Ablauf von Vorne mit einem neuen Frame.

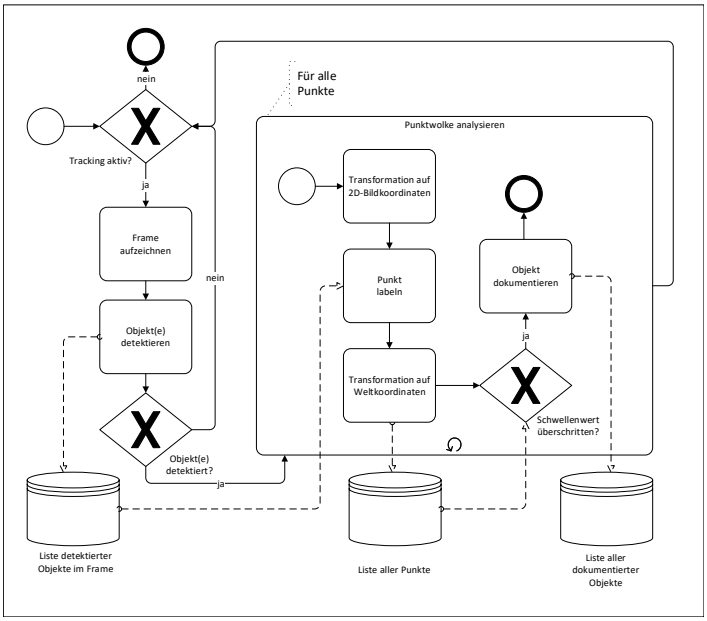



**Abbildung 2:** Prozessmodell des entwickelten Verfahrens

## 4.2 Testszenario

Zur Demonstration des Ansatzes wurde ein Raum gescannt, der über folgende detektierbaren Objekte verfügt: Stuhl, Tisch, TV, Couch, Flasche, Laptop, Zimmerpflanze, Fernbedienung und Computermaus. Tabelle 1 gibt einen Überblick über die Anzahl der Objekte. Weiterhin gibt die Spalte “Detektiert” Auskunft darüber, wieviele Objekte des jeweiligen Typs erkannt wurden. Die Spalte “Position” gibt an, wieviele Objekte an der richtigen Position (Lagegenauigkeit +/- 10cm) detektiert wurden. Die Farbangaben in Spalte fünf korrespondieren mit den Farben in Abbildung 3 Das Scannen des Raums dauerte ca. 1 Minute. Abbildung 3 a) visualisiert die dünnbesetzte Punktwolke, die im Rahmen der Positionsverfolgung durch ARCore generiert wurde. Mittels Objekterkennung gelabelte Punkte werden farblich differenziert nach Objekttyp dargestellt. Grafik 3 b) veranschaulicht das Ergebnis im Kontext einer Mesh-Darstellung des Raums. Das Mesh wurde mit einem Asus Zenfone AR unter Verwendung der Google-Tango-Technologie generiert.

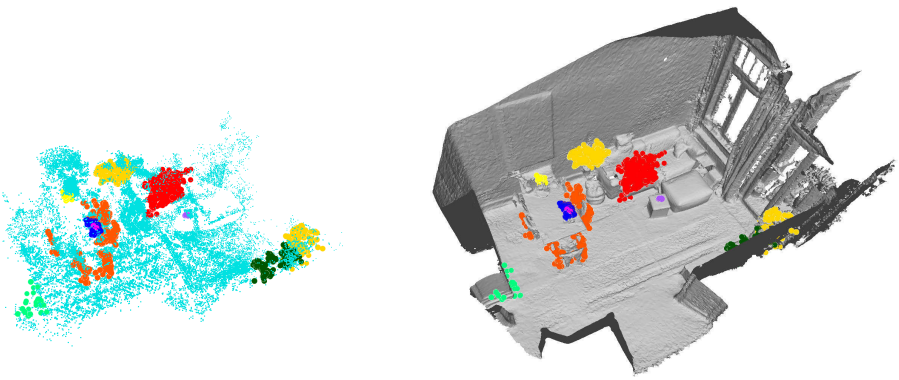
**Tabelle 1:** Auswertung Testszenario

| Objekt | Anzahl | Detektiert | Position | Farbe |
|---|---|---|---|---|
| Stuhl | 5 | 4 | 4 | orange |
| Tisch | 1 | 0 | 0 | braun |
| TV | 1 | 1 | 1 | dunkelgrün |
| Couch | 1 | 1 | 1 | rot |
| Flasche | 4 | 3 | 3 | hellgelb |

*wird auf der nächsten Seite fortgesetzt*

**Tabelle 1:** Auswertung Testszenario

| Objekt | Anzahl | Detektiert | Position | Farbe |
|---|---|---|---|---|
| Laptop | 1 | 1 | 1 | blau |
| Zimmerpflanze | 2 | 2 | 2 | dunkelgelb |
| Computermaus | 1 | 1 | 0 | pink |
| Schrank | 1 | 0 | 0 | hellrot |
| Kühlschrank | 0 | 1 | 0 | hellgrün |
| Fernbedienung | 2 | 2 | 2 | lila |



(a) (b)

**Abbildung 3:** a) gelabelte, dünnbesetzte Punktwolke der Positionsbestimmung nach Objekterkennung, b) Überblendung des Ergebnisses des konzipierten Verfahrens mit einem Mesh des Raums

## 5 Fazit und Ausblick

Die durchgeführten Tests haben gezeigt, dass das konzipierte Verfahren technisch realisierbar ist sowie autark auf handelsüblichen Geräten lauffähig ist und somit für (Tatort-)Dokumentationen verwendet werden könnte. Ein Problem, das aktuell einer praktischen Verwertung noch im Wege steht, ist die auftretende Falschklassifizierung von Objekten. Im oben gezeigten Testszenario wurde beispielsweise ein weißer Schrank als Kühlschrank klassifiziert. Auch war der Testraum sehr übersichtlich, für eine Verwendung in der Praxis müssen noch weitere Tests in ungeordneteren Umgebungen durchgeführt werden. Erweitert werden könnte der Ansatz auch dahingehend, dass für ein Verbrechen typische Objekte wie z.B. (Schuss-)Waffen dem neuronalen Netz angelernt werden. Diese Arbeit hat demonstriert, dass mit Hilfe von KI-Methoden Objekte automatisch in Echtzeit mit handelsüblichen Smartphones sowie Tablets autark am Einsatzort detektiert und verortet werden können. Im Rahmen von Tatortdokumentationen entsteht dadurch die Möglichkeit der automatischen Anreicherung eines räumlich-semantischen Tatortmodells. Dies reduziert den Aufwand der manuellen Datenaufnahme für die durchführenden Personen und gestaltet die gesamte Dokumentation effizienter.

## Literatur

Apple (2019), 'ARKit | Apple Developer Documentation', https://developer.apple.com/documentation/arkit.

Beetz, J., Blümel, I., Dietze, S., Fetahui, B., Gadiraju, U., Hecher, M., Krijnen, T., Lindlar, M., Tamke, M., Wessel, R. and Yu, R. (2016), Enrichment and Preservation of Architectural Knowledge, *in* '3D Research Challenges in Cultural Heritage II', Lecture Notes in Computer Science, Springer, Cham, pp. 231–255.

Czerniawski, T., Nahangi, M., Haas, C. and Walbridge, S. (2016), 'Pipe spool recognition in cluttered point clouds using a curvature-based shape descriptor', *Automation in Construction* **71**, 346–358.

Dai, J., Li, Y., He, K. and Sun, J. (2016), R-fcn: Object detection via region-based fully convolutional networks.

Franz, S., Irmler, R. and Rüppel, U. (2018), 'Real-time collaborative reconstruction of digital building models with mobile devices', *Advanced Engineering Informatics* **38**, 569–580.

Google (2019*a*), 'Google ARCore', https://developers.google.com/ar/reference/java/.

Google (2019*b*), 'TensorFlow examples', https://github.com/tensorflow/examples.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. (2014), 'Microsoft COCO: Common Objects in Context'.

Liu, L., Li, H. and Gruteser, M. (2019), 'Edge Assisted Real-time Object Detection for Mobile Augmented Reality', p. 16.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. (2016), SSD: Single Shot MultiBox Detector.

Qi, C. R., Litany, O., He, K. and Guibas, L. J. (2019), 'Deep Hough Voting for 3D Object Detection in Point Clouds', *arXiv:1904.09664 [cs]* .

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016), You Only Look Once: Unified, Real-Time Object Detection, *in* '2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, Las Vegas, NV, USA, pp. 779–788.

Ren, S., He, K., Girshick, R. and Sun, J. (2017), 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149.

Wald, J., Tateno, K., Sturm, J., Navab, N. and Tombari, F. (2018), 'Real-Time Fully Incremental Scene Understanding on Mobile Platforms', *IEEE Robotics and Automation Letters* **3**(4), 3402–3409.

Zhang, W., Han, B. and Hui, P. (2018), Jaguar: Low Latency Mobile Augmented Reality with Flexible Tracking, *in* '2018 ACM Multimedia Conference on Multimedia Conference - MM '18', ACM Press, Seoul, Republic of Korea, pp. 355–363.